

March 19, 1964

Dr. Emil L. Smith
Department of Biochemistry
University of California School of Medicine
Los Angeles, California

Dear Emil:

I would like to repeat how very much I enjoyed your lecture on the evolution of the cytochromes a couple of weeks ago, and particularly your thoughtful rebbttals to a couple of my spontaneous questions. The conservation of the cytochrome sequence between yeast and man is absolutely staggering and it has helped me to look at the issues of protein evolution in a completely different light, putting most of the stress on the functional requirements of the sequence.

I thought you might be interested in some versions of the "Torah" a little more compact than the rolls you displayed at the meeting, which I have found very helpful in looking at polypeptide sequences. These are off some of the computer programs we have been playing with lately in a so far not very fruitful analysis of common information in amino acid sequences. The enclosures on the cytochromes show the amino acids in Sorm's code, and with the amino acids separated into I hope not entirely arbitrary groups, the dissection helping one to see how these are organized along a sequence.

The Xerox copies also illustrate another matching program wherein the computer looks for the best match of an indicated pair of sequences with respect to the number of spaces that one sequence has to be shifted relative to another, and showing how far away one has to look to find a matching element in the opposite string. So, for example, the consistency of matching elements on line 5 for human heart cytochrome versus baker's yeast cytochrome is an indication of what you already know very well, the very close correspondence between these two strings subject to an extra set of five elements at the initial of the yeast. The number at line -20 refers to the position that the computer found most expeditious to initiate the search from at each level. The other similar plot of human heart versus pseudomonas cytochrome also shows what you very well know, namely the almost complete lack of homology between these two sequences, and the variety of numbers seen at -20 shows how far the computer program had to wander in finding the best available match over any appreciable interval, as poor as that was. These programs are organized to allow matching to be done under a variety of conditions of equivalents; the present display shows the condition of identify. We have spent some time in looking at other

corrections made on original

EMIL

SMITH

cc

more relaxed conditions, for example, functional equivalences similar to those shown on the dissected plots, as well as various types of lists of permitted mutations on either a theoretical or an experimental basis. However, I must say that these relaxed conditions of equivalences have not shown anything especially interesting except that the functional similarity does bring out very well the string of nine elements that begins at position 80 in human heart cytochrome, MWFVYWIII, which is a pretty good match to the sequence NLFYYLII in the baker's yeast cytochrome, and which also has a corresponding sequence of nine elements in myoglobin. It is also pretty plain that the string of hydrophobic elements followed by the polar amino acids is relatively relaxed with regard to which polar amino acid follows precisely which, just so long as the total charge in that region roughly balances out. At least, this is the impression that I get comparing the myoglobins and the cytochromes together. But I certainly would not suggest that there is a necessary phylogenetic relationship between these elements of structure since it is such an obvious building block for a segment of protein to perform God knows what function. I have to state too that in a test of randomized sequences we did find a match just as good as this in just one out of twenty comparisons of the same protein strings, so it is difficult to attribute even a statistical significance to this, though I find it hard to avoid that there is some sense locked up in this group of nine elements. Since we have reached this impasse, it is not plain that there is very much more constructive that we can do with these amino acid comparison programs, but I would be delighted to have any suggestions from you. As more and more information does come out on sequences of a variety of proteins, I am quite confident that this approach will be more and more essential. ce

Cordially, and with the best of luck (perhaps a little belatedly now) for your new post,

Joshua Lederberg
Professor of Genetics